Annotation des descriptions définies : le cas des reprises par les rôles thématiques

Hélène Manuélian

LORIA
Campus Scientifique
BP 239
54 506 Vandoeuvre Lès Nancy
helene.manuelian@loria.fr
Fin de thèse : décembre 2003

Mots-clefs – Keywords

Génération automatique de textes, annotation, coréférence, rôles thématiques

Natural language generation, annotation, coreference, thematic role

Résumé – Abstract

Nous présentons dans cet article un cas particulier de description définie où la description reprend le rôle thématique d'un argument (implicite ou explicite) d'un événement mentionné dans le contexte linguistique. Nous commençons par montrer que les schémas d'annotation proposés (MATE) et utilisés (Poesio et Vieira 2000) ne permettent pas une caractérisation uniforme ni, partant, un repérage facile de ces reprises. Nous proposons une extension du schéma MATE qui pallie cette difficulté.

We present in this paper a particular kind of definite description which mentions the thematic role of an (implicit or explicit) argument of an event mentioned in the context. We show that the annotation schemes which are either proposed (MATE) or used (Poesio and Vieira 2000) permit neither a uniform characterisation of the phenomenon nor an easy identification of these anaphors. We propose an extension of the MATE schemes which remedy this difficulty.

1 Introduction

Le travail présenté dans cet article se situe dans le cadre de la génération automatique d'expressions référentielles. Nous cherchons à générer des textes contenant des reprises anaphoriques par la verbalisation de rôles thématiques :

> *Cinq millions d'Euros* ***ont été versés*** *à Amnesty International.* ***Le généreux donateur*** *préfère garder l'anonymat.*

Notre but est d'arriver à une extension de l'algorithme de Dale et Reiter (1995) par une prise en compte plus importante du contexte linguistique. En effet, l'algorithme de génération

d'expressions référentielles mis au point par Dale et Reiter permet de référer à un objet grâce à ses propriétés distinctives dans le contexte de référence mais ne tient pas compte de la totalité du contexte linguistique. Il exige l'unicité du référent dans le contexte mais nous verrons (section 2) que si cette condition est nécessaire, elle n'est pas toujours suffisante pour la génération d'une description définie. L'algorithme pourra donc générer indifféremment les textes suivants :

> *L'ETA **a** de nouveau **tué** hier. **La victime** est un député conservateur.*
>
> *? L'ETA **a** de nouveau **tué** hier. **L'homme politique** est un député conservateur.*[1]

Même si le groupe nominal permet d'identifier le personnage dans le contexte, le second texte est beaucoup moins cohérent que le premier. Parvenir à générer une reprise par le rôle thématique est donc un bon moyen d'assurer la cohérence du texte.

Avant de proposer une extension d'algorithme pour ce phénomène, nous avons fait des recherches sur des corpus annotés afin d'en étudier la fréquence et la distribution, et d'établir des contraintes sur sa génération. Ces recherches nous ont amenée à dire que les schémas d'annotation existant ne sont pas suffisants pour l'isoler, et à proposer une extension de ces schémas. Dans les sections 2 et 3, nous présentons le phénomène étudié et nous nous positionnons par rapport aux travaux sur le sujet en linguistique. Dans la section 4, nous présentons notre étude du corpus annoté au niveau référentiel par Poesio et Vieira (2000). Nous montrons en quoi les classifications des utilisations des descriptions définies et les schémas d'annotation qui y sont proposés sont insuffisants. Enfin, dans la section 5, nous présentons le schéma d'annotation proposé par le projet MATE[2] pour ce phénomène, ses insuffisances et notre proposition d'extension de ce schéma.

2 Utilisations des descriptions définies

On considère traditionnellement que le déterminant défini donne deux indications sur l'entité dénotée par le syntagme nominal : elle est unique dans le contexte (Russell 1905, Corblin, 1987) et elle est connue (Heim 1982). Il existe un ensemble de nuances à cette observation, qui amènent Vieira (1998) à construire en quatre grandes catégories d'utilisation du défini que nous décrivons brièvement ici :

- **Utilisation anaphorique.** Le groupe nominal entretient une relation de coréférence avec un autre élément linguistique présent dans le contexte linguistique antérieur. Les deux segments linguistiques réfèrent à la même entité, peuvent être identiques ou différents. (deux noms identiques, deux noms différents...)

- **Utilisation associative.** Le groupe nominal entretient une relation avec une expression qui ne réfère pas exactement à la même entité, mais les connaissances du monde permettent de faire un lien très facilement entre les deux entités dénotées par les deux expressions, tant leur relation est forte. (Ces relations sont la plupart du temps des relations méréologiques – d'une partie vers un tout ou d'un élément à un ensemble.) La présence de la première entité dans le contexte implique presque automatiquement la présence de la seconde. Lorsqu'il s'agit de deux groupes nominaux, les têtes de syntagmes sont automatiquement différentes, mais il peut parfois s'agir d'une relation

---

[1] Nous utiliserons la notation suivante : les textes dont la compréhension nous semble douteuse seront précédés de ? et ceux qui nous semblent agrammaticaux seront précédés de *.

[2] Multilevel Annotation Tools Engineering, Telematics Project LE4-8370.

entre un groupe verbal et un groupe nominal, le groupe nominal référant à l'un des participants de l'action qui n'est pas mentionné explicitement.

- **Utilisation situationnelle.** Les utilisations situationnelles sont possibles parce que l'entité à laquelle elles réfèrent est présente dans le contexte général du discours ou dans la situation des interlocuteurs au moment de la production de l'énoncé.

- **Utilisation non familière.** Ces utilisations sont quasiment toujours des premières mentions. Elles réfèrent à des entités uniques, qui ne sont pas forcément connues de l'interlocuteur. En général, le nom et ses modifieurs forment une définition suffisante de l'objet pour qu'il soit identifié comme unique dans le contexte, bien qu'il n'ait pas été mentionné antérieurement.

3 Reprises par le nom du rôle thématique

Nous nous basons sur les travaux de Jackendoff (1990), qui représente le sens des mots par des représentations mentales appelées structures lexicales conceptuelles (LCS). Dans la LCS, sont définies des positions pour les arguments sémantiques du verbe qui correspondent la plupart du temps aux rôles thématiques qu'il sélectionne. Ainsi, le verbe *aller* sélectionne une destination (position TO), un point de départ (position FROM), et un élément effectuant le déplacement (THEME). L'ensemble de ces deux positions forme un troisième argument de type chemin (position VIA). Aussi, quand nous parlerons de rôles thématiques ici, il s'agira précisément de ces positions définies dans la LCS. Les positions sont considérées comme des fonctions, et reçoivent des types issus d'une ontologie relativement simple. Le verbe aller (GO, sans type) au sens « aller d'un lieu à un autre » aura la même structure que le verbe donner « aller d'un possesseur à un autre » (GO avec le type *poss*), mais les positions recevront un type ontologique différent, ce qui permettra de distinguer ces verbes :

$$\text{GO}\left[\text{THEME}[_{\text{thing}}], \text{VIA}\left[\begin{array}{c}\text{FROM}[_{\text{place}}]\\ \text{TO}[_{\text{place}}]\end{array}\right]_{\text{path}}\right] \quad \text{GO}_{\text{poss}}\left[\text{THEME}[_{\text{thing}}], \text{VIA}\left[\begin{array}{c}\text{FROM}[_{\text{thing}}]\\ \text{TO}[_{\text{thing}}]\end{array}\right]_{\text{path}}\right]$$

Nous voyons donc que les rôles thématiques sont en fait l'application de fonctions sur des positions dans la structure argumentale des verbes. Les types des positions, permettent d'avoir des structures très générales, tout en conservant le maximum d'information concernant les restrictions sélectionnelles. La lexicalisation de ces positions est le résultat de l'application d'une fonction sur un type ontologique. Ainsi, la position TO dans la LCS de *aller* sera lexicalisée par le terme *destination*, tandis que dans la LCS du verbe *donner* elle sera lexicalisée par *bénéficiaire*. Nous montrons dans cette section comment les lexicalisations de ces positions dans la LCS peuvent permettre de référer au participant d'un événement dans une description définie.

3.1 Reprises par le nom du rôle thématique d'un argument implicite

Parmi les utilisations associatives listées par Vieira (1998), on trouve des emplois où le syntagme nominal défini réfère à un argument du verbe de la phrase précédente, sans qu'il ait été mentionné. Le nom tête du syntagme sera alors la verbalisation du rôle thématique joué par l'entité dans l'événement précédemment mentionné. Il s'agit de la catégorie que Clark (1977) nomme « référence indirecte par caractérisation », dont voici deux exemples :

1) *Cinq millions d'Euros **ont été versés** à Amnesty International. **Le généreux donateur** préfère garder l'anonymat.*

2) *Bob **est allé** à Venise**. La route** a été pénible.*

Dans l'exemple 1, bien que l'agent effectuant le don ne soit pas mentionné dans la première phrase, on y fait référence dans la seconde, grâce à une description définie verbalisant son rôle dans l'action (position FROM avec donner). Dans l'exemple 2, on réfère à la totalité du déplacement par *la route*, alors que seul le point d'arrivée est mentionné dans la première phrase. On peut dire qu'il s'agit de la verbalisation de l'application de la fonction VIA à ses deux arguments dans les LCS de Jackendoff. Dans la classification de Clark, ces deux exemples appartiennent à la catégorie des références indirectes par caractérisation d'un argument obligatoire. (On ne peut pas envisager l'événement correspondant à verser de l'argent sans un agent effectuant cette action, ni un déplacement sans chemin.)

L'exemple qui suit est classé par Clark dans la catégorie de la référence indirecte par caractérisation d'un argument optionnel, un attentat ne faisant pas forcément de victime :

*3) L'ETA **a commis un attentat** hier soir. **La victime** est un député conservateur.*

Nous considérons ici *commettre un attentat* comme une collocation, ayant pour arguments sémantiques un agent, une cible, et un instrument. Un des arguments sémantiques de *commettre-un-attentat* n'est pas réalisé dans la première phrase. Un attentat peut impliquer une victime, c'est pourquoi on peut se permettre d'utiliser un groupe nominal défini dans la seconde phrase du texte, bien que le personnage ne soit pas mentionné précédemment.

Enfin, nous pouvons trouver le même type de phénomène entre deux groupes nominaux, lorsque le premier nom fait référence à un événement :

*4) L'ETA revendique un nouvel **attentat**. **La victime** est un journaliste d'El Paìs.*

Ici, *la victime* réfère à un argument du prédicat *attentat* (on peut d'ailleurs écrire *la victime de l'attentat*), de la même manière que dans l'exemple 3. Il s'agit d'un cas spécifique d'anaphore associative où la reprise se fait par le rôle thématique. (La plupart du temps on parle d'anaphore associative quand elle a lieu par le biais d'une relation méréologique)

3.2 Reprise par le nom du rôle thématique d'un argument mentionné

*5) **Le PDG de Total-Fina-Elf** a été assassiné hier soir. **La victime** se rendait au tribunal quand deux motards ont surgi et lui ont tiré dessus.*

Dans cet exemple, l'entité à laquelle on réfère par un nom de rôle thématique a déjà été mentionnée. Cependant, la propriété « victime » du « PDG de Total-Fina-Elf » n'est pas intrinsèque au personnage, mais introduite par le verbe *assassiner*. Si on changeait de contexte, même après le décès du personnage, on ne pourrait pas le décrire ainsi :

*6) *Le PDG a mal dormi cette nuit. La victime s'est rendue à son bureau très tôt ce matin.*

L'interprétation et la génération de *la victime* nécessitent un appui sur le premier SN ***et*** sur le verbe.

Dans tous les cas, on constate la lexicalisation d'un rôle thématique dans la deuxième phrase. Dans les exemples 1 à 4, il s'agit de la lexicalisation d'une entité non mentionnée explicitement dans la description de l'événement, mais présente dans sa structure sémantique. Dans le cinquième exemple, l'entité a déjà été mentionnée dans la première phrase, mais la reprise se fait par le nom correspondant au patient de l'événement « assassinat ». Dans tous les cas, l'antécédent est soit un verbe, soit un nom déverbal.

3.3 Les reprises par les noms de rôles thématiques vues comme des cas particuliers d'anaphore associative

Nous considérons comme Clark (1977) que les cas décrit en section 3.1 est sont des anaphores associatives dont les antécédents sont des verbes ou des noms déverbaux et nous le démontrons par rapport à la classification établie dans la section précédente :
(i) Il ne s'agit pas d'une utilisation anaphorique (section 2.1) puisque l'entité à laquelle on réfère n'est pas mentionnée précédemment dans le texte.
(ii) Il ne s'agit pas non plus d'une nouvelle entité dans le discours (utilisations non familières, section 2.4), puisque l'interprétation du syntagme ne serait pas possible dans un autre contexte linguistique (cf. exemple 6). En effet, lors d'une première mention, on utilise une propriété intrinsèque de l'entité qui la différencie de toutes les autres entités du contexte.
(iii) Enfin, il ne s'agit pas d'une anaphore situationnelle (section 2.3), comme dans « le Soleil » ou « le Premier Ministre », puisqu'il ne s'agit pas de faire référence à un objet présent dans la situation d'énonciation, mais bien dans le contexte discursif.

En revanche, toutes les caractéristiques de l'anaphore associative s'appliquent à nos exemples : l'entité est implicitement présente dans le contexte grâce à la mention d'un autre élément dans le contexte antérieur, et on ne peut interpréter le syntagme qu'en lien avec cet élément linguistique. Les cas que nous étudions diffèrent des cas les plus fréquemment étudiés dans la littérature par le fait que la relation entre l'antécédent et l'anaphore est thématique et non pas méréologique.

4 Etude de corpus

Dans cette section, nous présentons notre travail sur corpus. Dans un premier temps nous avons mené des recherches informelles sur un corpus français pour vérifier l'existence du phénomène. Nous avons ensuite fait des recherches plus formelles sur un important corpus annoté pour les descriptions définies, dans le but de vérifier la pertinence du schéma d'annotation utilisé dans la pratique pour les descriptions définies.

4.1 Exploitation informelle du corpus du Monde Diplomatique

Nous utilisons un corpus rassemblant les articles parus dans *Le Monde Diplomatique* en 1996 et 1998[3], choisi car le style journalistique emploie fréquemment le type de reprise anaphorique que nous recherchons. Ce corpus n'est pas annoté, nous avons donc dû définir une stratégie de recherche informelle d'exemples de reprise par le nom d'un rôle thématique.

En nous basant sur Jackendoff (1990), nous avons recherché une liste de verbes instanciant un schéma du type GO – FROM – TO – VIA, c'est à dire des verbes qui décrivent le passage d'un objet d'un lieu à un autre, d'un possesseur à un autre ou d'un état à un autre. Nous avons choisi les verbes *aller, acquérir* et *mourir* ainsi que des verbes synonymes, (*arriver, se diriger (vers), voyager, visiter, recevoir, gagner*), des verbes ayant un sens converse (*quitter, partir, prendre, acheter*), et leurs équivalents avec un sens causatif (*conduire, emmener, donner, vendre, tuer, assassiner*)

Nous avons ensuite cherché comment on pouvait lexicaliser les rôles correspondant aux arguments sémantiques FROM, TO, VIA et THEME que ces verbes sélectionnent (*destination, provenance, voyageur, trajet, destinataire, bénéficiaire, don, donneur, donateur, victime, etc.*) et

[3] Corpus SILFIDE, Equipe Langue et Dialogue du LORIA, http://www.loria.fr/projets/Silfide/. Le corpus Monde diplomatique 96 contient 939 489 mots et le Monde diplomatique 98 en contient 565 900.

cherché leurs synonymes dans le dictionnaire de synonymes en ligne de l'InaLF complété par le CRISCO (Caen) (http://elsap1.unicaen.fr/dicosyn.html).

Nous avons enfin effectué des recherches sur ces noms de rôle dans les deux années du *Monde Diplomatique*, et repéré des exemples attestant de l'existence de ce phénomène[4].

> *Depuis, le clan de Donetsk **a été décimé par une série de meurtres**, **dont le plus marquant** a été commis le 3 novembre 1996 : **la victime**, Yevhen Shcherban, était un homme d'affaires important et chef présumé du clan. (File "Lemonde98.sgml"; Line 30449)*
>
> *Des fondations reposant sur de grandes et anciennes fortunes industrielles américaines(...), **financent** aussi des chaires dans les universités les plus prestigieuses des Etats-Unis(...) la Fondation Olin **qui consacrait déjà, en 1988, 55 millions de dollar**s à cet objectif. Il va de soi qu'avec des sommes pareilles **le généreux donateur** a le droit de nommer les professeurs qui vont occuper les chaires et diriger les centres d'études (14). (File "Lemonde96.sgml"; Line 57154)*
>
> *Les voisins dormaient déjà lorsqu'ils **arrivèrent**. Elle réveilla son enfant qui avait dormi à poings fermés pendant tout **le voyage**. (File "Lemonde96.sgml"; Line 1103)*

4.2 Exploitation formelle du corpus de Poesio et Vieira (2000)

A la suite de nos recherches sur le *Monde Diplomatique*, nous souhaitions vérifier nos résultats en faisant une recherche plus systématique et approfondie des occurrences de ce phénomène. Nous avons donc consulté un corpus annoté : celui dont nous disposons est le corpus annoté par Poesio et Vieira (2000). Il s'agit d'un sous ensemble du corpus Penn Treebank I, qui est l'un des corpus les plus importants annoté au niveau référentiel (1000 descriptions définies annotées).

4.2.1 Schéma d'annotation utilisé

Dans ce corpus, quatre catégories de description définies ont été annotées :

- **Anaphore directe** : Il s'agit des références subséquentes à un objet utilisant la même tête nominale dans les deux syntagmes : *Le petit chat noir est entré dans la maison… le chat…*
- **Associations** : Il s'agit de deux types de relation : (a) mention d'une nouvelle entité en lien avec une entité mentionnée précédemment dans le discours (anaphore associative) : *Au loin, on aperçoit une maison. La porte est entrouverte.* (b) mention de la même entité avec un nom différent comme tête du syntagme (coréférence) : *Au loin, on aperçoit une maison. Le bâtiment…*
- **Première mention** : Il s'agit des descriptions référant à un objet sans lien avec le contexte linguistique ou général.
- **Non résolu** : Cette catégorie regroupe toutes les descriptions n'entrant pas dans les autres catégories.

---

[4] Sans pour autant pouvoir faire des statistiques fiables, dans la mesure où la recherche n'était pas automatique.

Etant donné ce schéma d'annotation, on aurait pu s'attendre à trouver des exemples de reprises par un rôle thématique implicite dans la catégorie des *premières mentions*, et les reprises par un rôle thématique explicite dans la catégorie *associations*, annotées en coréférence avec le premier syntagme nominal référant à l'objet dans le texte. Nous avons donc étudié ces catégories en recherchant tous les noms correspondant à des rôles dans des événements. Nous n'avons trouvé aucune occurrence de ce phénomène classé dans ces deux catégories. En revanche nous en avons trouvé dans la catégorie *Non-résolu* :

*7) So the IRS has drawn a rationale from the **sale of a home site split in two** and sold in different years to ** **the same buyer** **.*
*<< the same buyer >> *** UNRESOLVED *** (File "Sample"; Line 8707)*

*8) The changes were proposed in an effort to streamline federal bureaucracy and boost compliance by the executives who are really calling the shots, said Brian Lane, special counsel at the SEC's office of disclosure policy, which proposed **the changes**. Investors , money managers and corporate officials had until today to comment on ** **the proposals** ** , and the issue has produced more mail than almost any other issue in memory , Mr. Lane said.*
*<< the proposals >> *** UNRESOLVED ***(File "Sample"; Line 294)*

Dans l'exemple 7, le groupe nominal *the same buyer* est considéré comme non résolu, alors qu'il s'agit de la lexicalisation de la position TO dans la structure lexicale conceptuelle de *sale* dans le groupe nominal *the sale of a home split in two* (cf. exemple 5). Dans l'exemple 8, la production du syntagme *the proposals* dans la deuxième phrase n'est possible que grâce au verbe *proposed* dans la première phrase. Il s'agit d'une reprise du thème du verbe *to propose*, réalisé explicitement dans la première phrase par *the changes* (cf. exemple 4).

Nous pensons que les reprises anaphoriques recherchées n'ont pas été identifiées par le système pour deux raisons :

Premièrement, les définitions des catégories d'emploi des définis ne sont pas assez précises. D'une part, il n'est pas explicite que l'antécédent d'une anaphore puisse aussi être verbal, et d'autre part, il n'y a pas d'indication dans la catégorie *associations* expliquant que la reprise peut se faire par un nom de rôle thématique. On hésite alors à faire coréférer deux noms dont les référents n'entretiennent pas de relation méréologique, dans la mesure où aucun exemple d'un autre type n'est donné dans les instructions aux annotateurs.

Par ailleurs, si des heuristiques ont été construites pour retrouver certains cas d'anaphore associative, aucune n'a été mise au point pour le type de reprise que nous cherchons, ce qui classe automatiquement les exemples que nous recherchons dans la catégorie *non-résolu*.

En pratique et en théorie, le schéma d'annotation semble mal adapté à une caractérisation des reprises par le rôle thématique. Sur le plan pratique, on a vu qu'il n'est pas suffisamment précis pour permettre une annotation adéquate (les reprises que nous cherchons sont dans la catégorie *non-résolu*). Sur le plan théorique, le schéma ne permet pas une caractérisation uniforme et par conséquent un repérage facile des reprises thématiques (le fait que le nom reprend un rôle thématique n'est pas annoté).

5 Annotation proposée

Les recommandations MATE (Davies et al. 1998) ont été écrites de façon à aller vers un schéma d'annotation de la coréférence uniforme pour tous les corpus. Les reprises anaphoriques sont annotées par une balise <link> qui sera typée. Les recommandations distinguent – entre autres – trois types de liens anaphoriques. Tout d'abord les groupes nominaux coréférents à un

argument mentionné explicitement seront annotés par un type *ident*. Dans les cas d'anaphore associative, le lien sera de type *element.* Enfin dans les cas de reprise par un nom de rôle thématique, le type de la balise <link> sera *e-rel*. On trouve une autre balise importante dans la balise <link >: la balise *arg*, dont la valeur est une paire composée de l'identifiant de l'antécédent et de l'identifiant du groupe nominal annoté. Dans les deux premiers cas (*ident* et *element*), l'antécédent sera un groupe nominal, et dans le troisième cas, l'antécédent sera la proposition complète à partir de laquelle la reprise sur le rôle thématique a lieu.

Le schéma pose les difficultés suivantes : d'une part, comment un annotateur va-t-il choisir entre une coréférence avec un syntagme nominal ou une coréférence avec la totalité de la proposition dans le cas de l'exemple 4 ? La conséquence de cette ambiguïté peut conduire à des désaccords entre annotateurs et donc à un indice kappa[5] faible et à des difficultés dans l'automatisation du système. D'autre part, cette annotation ne permet pas un repérage systématique du phénomène qui nous intéresse dans la mesure où (i) les reprises par le rôle thématique ne figurent pas systématiquement dans la même catégorie, (ii) on n'aura pas d'indication sur le rôle thématique si la description est classée dans la catégorie des coréférences, (iii) on n'aura pas non plus cette indication si elle est classée dans la catégorie des anaphores associatives.

Pour pallier ces insuffisances, nous proposons le schéma d'annotation suivant :

[5] indice d'accord entre les annotateurs. Le test kappa permet de calculer un coefficient mesurant la cohérence des annotations lorsqu'elles sont réalisées par plusieurs personnes, Carletta (1996).

Type de reprise	Balises < link >	Exemple
SN entretenant une relation anaphorique avec un argument implicite du verbe de la phrase précédente	< link type = “e-rel” role = “R” args = “X, Y”>	L’ETA <de ID = 1> **a commis un attentat** </de>. <de ID = 2> **La victime** </de> est un député conservateur. < link type = “e-rel” role = “Patient” args = “1, 2”>
SN entretenant une relation anaphorique avec une position sémantique de la LCS et une relation coréférentielle avec un argument du verbe.	< link type = “e-rel” role = “R” args = “X, Y”> < link type = “ident” args = “X, Z”>	<de ID = 1> Le PDG </de> <de ID = 2> **a été assassiné** </de>. <de ID = 3> **La victime** </de> se rendait à son bureau lorsqu’on lui a tiré dessus. < link type = “e-rel” role = “patient” args = “2, 3”> < link type = “ident” args = “1, 3”>
SN entretenant une relation anaphorique avec l’argument sémantique non réalisé d’une nominalisation d’événement	< link type = “element” role = “R” args = “X, Y”>	L’ETA revendique <de ID = 1> un nouvel **attentat** </de>.<de ID = 2> **La victime** </de> est un journaliste d’El Paìs. < link type = “element” role = “patient” args = “1, 2”>
SN coréférent à l’argument explicite d’une nominalisation d’événement, et anaphorique avec sa position dans la LCS.	< link type = “element” role = “R” args = “X, Y”> < link type = “ident” args = “X, Z”>	<de ID = 1> **Le meurtre** </de> de <de ID = 2> **Lennon** </de> a ému tout le pays.<de ID =3> **La victime** </de> était très populaire. < link type = “element” role = “patient” args = “1, 3”> < link type = “ident” args = “1, 2”>

Les différences entre nos propositions et le schéma MATE sont les suivantes :
Pour tous les cas de reprise par le rôle thématique, le rôle repris fait partie de l’annotation de la description définie grâce à la balise role = “R”. (La liste des rôles à utiliser reste à définir.)
Dans le cas des reprises d’un argument implicite de verbe ou de nom déverbal, la description pointe sur le nom/verbe antécédent.
Enfin, dans les cas de reprise d’un argument explicite, la description définie pointe sur le nom/verbe antécédent, ainsi que sur le groupe nominal coréférent.

6 Conclusion et perspectives

Partant des schémas d’annotation proposés (MATE) ou utilisés (Poesio et Vieira 2000), nous avons proposé une extension de MATE qui permet une caractérisation uniforme du phénomène de reprise par le rôle thématique. Pour permettre une analyse des contraintes linguistiques régissant ces reprises, nous voulons maintenant utiliser ce schéma pour une annotation systématique du corpus PAROLE[6], annotés au niveau morpho-syntaxique.

Nous souhaitons par ailleurs étendre notre étude au même phénomène, mais dans le cadre des descriptions démonstratives. Nous avons en effet trouvé des cas où il est nécessaire d’employer

[6] Corpus annoté au niveau morpho-syntaxique par l’ATILF. L’annotation se fait grâce à la collaboration de l’ATILF et du LORIA dans le cadre du plan Etat-Région. Nous supposons que ce corpus contiendra autant de reprises par le rôle thématique que le corpus SILFIDE, les textes étant de la même nature.

le démonstratif si on désigne l'entité par son rôle dans l'action, et il nous semble important de pouvoir déterminer des contraintes sur cet emploi du démonstratif et d'expliquer la différence entre les deux textes suivants :

**Bob est allé à Rome. La destination lui plaît beaucoup.*

Bob est allé à Rome. Cette destination lui plaît beaucoup.

Nous allons par ailleurs définir, en collaboration avec l'Université de Sarrebruck, un schéma d'annotation plus large des anaphores associatives en français et en anglais. Il nous semble nécessaire de quantifier le phénomène parmi les descriptions définies, et d'affiner notre schéma afin d'annoter le type de relation lexicale entretenues par l'antécédent et l'anaphore, en corrélant ces résultats avec le rôle thématique de l'entité et son type ontologique, ainsi que la fonction grammaticale du syntagme nominal. En effet, tous ces paramètre semblent se conjuguer pour parvenir au choix du déterminant, ou pour rendre possible l'anaphore associative. Partant, nous espérons parvenir à une analyse rigoureuse des corpus annotés et donner des bases empiriques et théoriques solides à l'extension des algorithmes de génération d'expressions référentielles, de manière à augmenter encore le caractère naturel des reprises anaphoriques dans les textes générés automatiquement.

Remerciements

Mille mercis à Claire Gardent pour son soutien et ses encouragements, ainsi qu'à Jean-Marie Pierrel. Merci à Renata Vieira, qui m'a fourni le corpus Penn Treebank I annoté, à Susanne Salmon-Alt et Josette Lecomte. Merci aussi à Evelyne, Fred et Djamé.

Références

Carletta J. (1996) Assessing Agreement on classification tasks : the kappa statistic, *Computational Linguistics*, 22(2) : 249-254

Clark H. H. (1977) *Bridging*. In P.N. Johnson-Laird and P.C. Wason, editors, *Thinking: Readings in Cognitive Science*. Cambridge University Press, London and New York.

Corblin F. (1987), *Indéfini, Défini et Démonstratif*, Droz, Genève – Paris.

Dale R., Reiter E. (1995) Computational Interpretation of the Gricean Maxims in the Generation of Referring Expressions, *Cognitive Science*, 19 : 233-263.

Davies S., Poesio M., Brunesaux F., Romary L., (1998) *Annotating Coreference in Dialogues : Proposal for a Scheme for MATE*, Report.

Heim I. (1982) *The Semantics of Definite and Indefinite Noun Phrases in English*. PhD Thesis, University of Massachusetts, Ahmerst.

Jackendoff R. (1990) *Semantic Structures*, MIT Press, Cambridge, Massachusetts, London, England.

Russell B. (1905) On denoting, *Mind*, 14 : 479-493.

Vieira R. (1998) A review of the Linguistic litterature on definite descriptions, *Acta Semiotica et Lingvistica*, Vol. 7 : 219-258.

Vieira R., Poesio M. (2000) An Empirically-Based System for Processing Definite Descriptions, *Computational Linguistics*, v. 26, n.4. 525-579.